Introduction to Functional Data Analysis

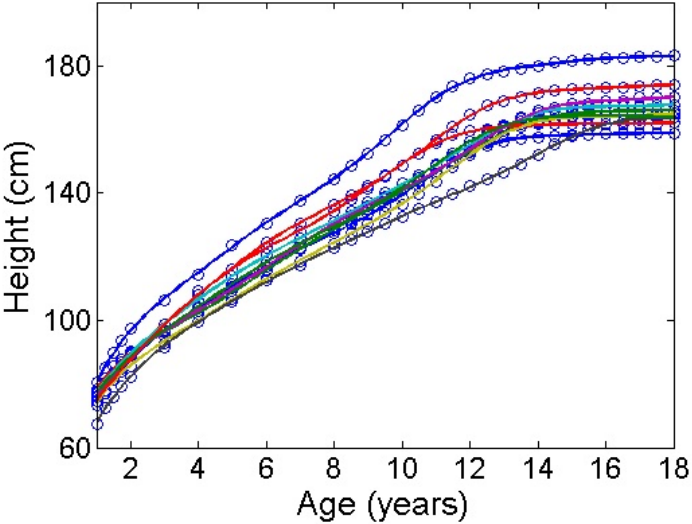
Pang Du

Department of Statistics
Virginia Tech

Overview

- What are functional data?
- Some functional data analyses
- Goals of functional data analysis
- First steps in functional data analysis
- Using derivatives in functional data analysis

Heights of ten girls



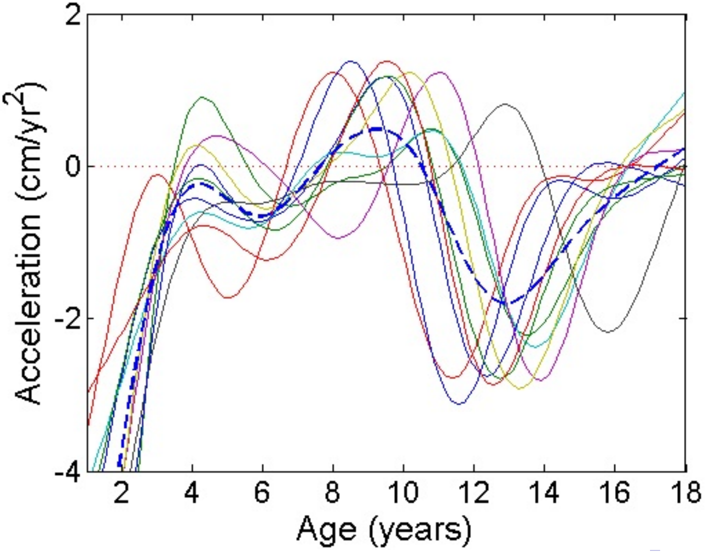
Data challenges

- We need repeated and regular access to each child for up to 20 years.
- Unequally-spaced measurement times: 4 in age of one, annual from age two to eight biannual afterwards.
- Height changes over the day and must be measured at a fixed time.
- Height is measured in supine position in infancy, followed by standing height. The change involves an adjustment of about 1 cm.
- Measurement error is about 0.5 cm in later years, but rather large in infancy.

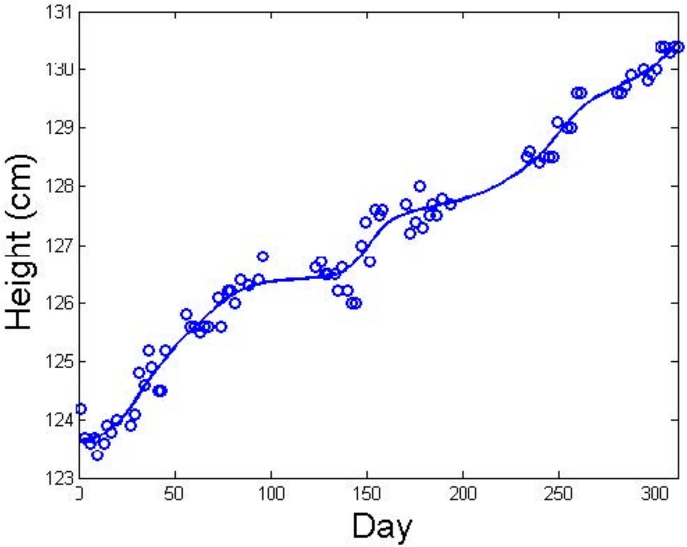
Modeling challenges

- We want *smooth* curves that fit the data as well as is reasonable, i.e., with a typical error level that starts at about 0.7 cm but decrease to around 0.5 cm.
- In principle the curves should be *monotone*.
- We also want to look at velocity and acceleration, that is, we need the first two *derivatives* of the height *function*.

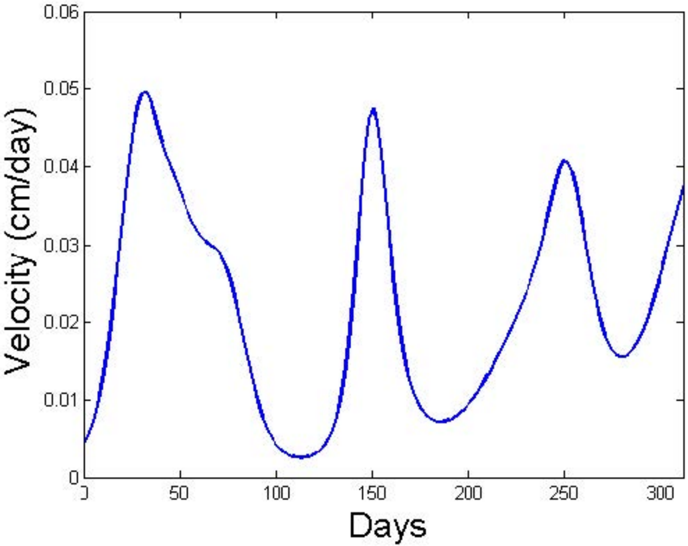
Height accelerations of ten girls



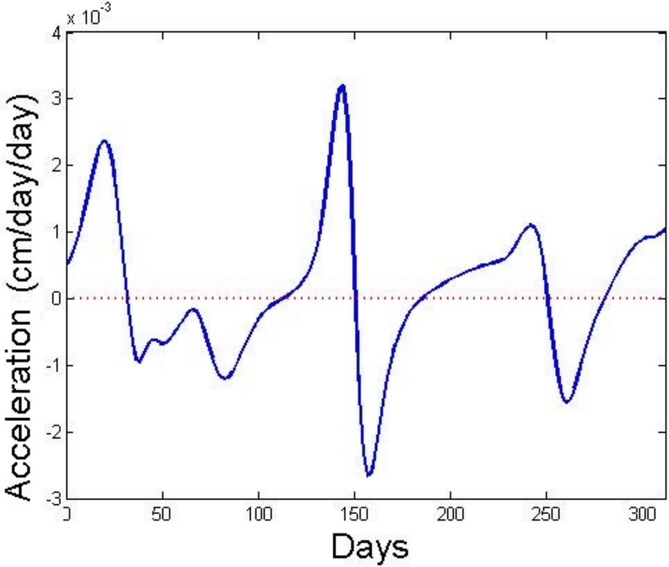
Height of a 10-year-old boy



Height velocity of the boy



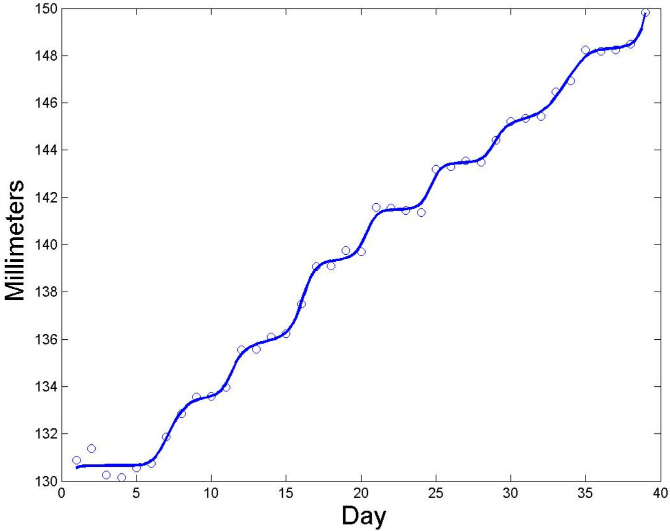
Height acceleration of the boy



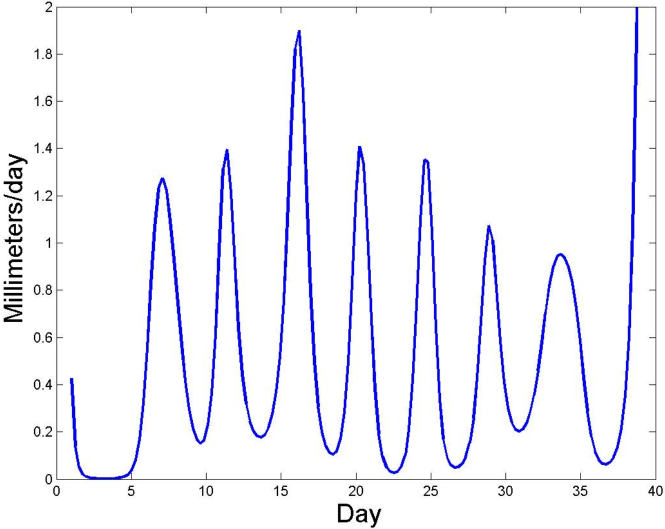
Tibia length of a newborn baby

- Prof. Michael Hermanussen (Kiel, Germany) developed an instrument capable of measuring the length of the tibia of a baby (the lower leg bone) with an accuracy of about 0.1 millimeters.
- He measured newborn infant's tibias daily and hourly.

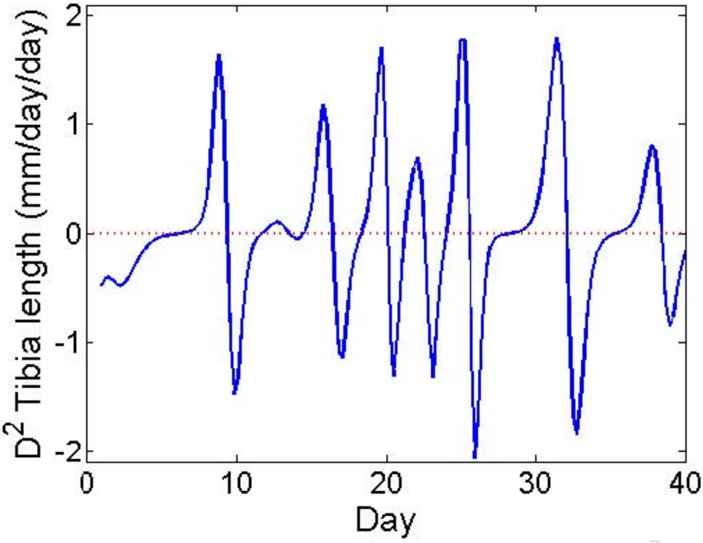
Height of a newborn baby



Height velocity of a newborn baby



Height acceleration of a newborn boy



Some conclusions about growth

- Over 20 years, there is one major growth spurt (pubertal growth spurt), but clear evidence for at least one minor spurt.
- The timing of these spurts varies from child to child.
- Zooming in on a daily scale, at ten years of age there is a growth spurt every 100 days or so, and the amount of energy in the spurts seems to be decreasing.
- A newborn's tibia can grow at an astonishing 0.5 millimeters per day!
- A critical aspect of growth is what shuts it off.

Issues for further investigation

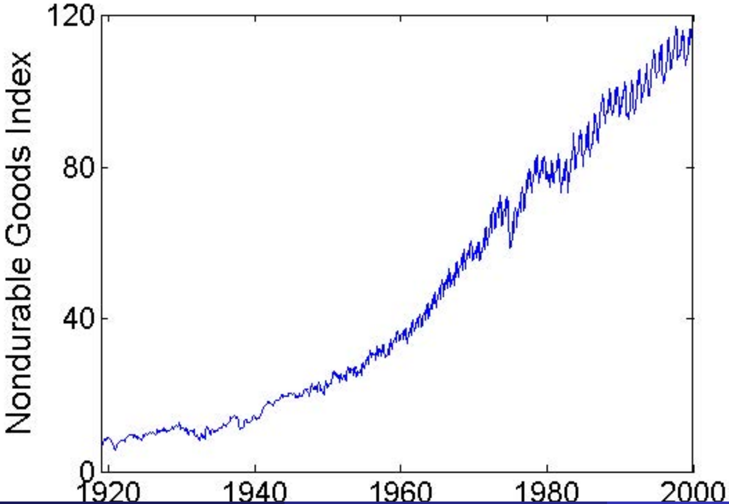
For the girls height study, further investigation of the variation across curves is of interest.

- Phase variation: different growth spurt timing.
- Amplitude variation: different intensities of growth.
- Separation of these two sources of variation is important: *curve registration*.

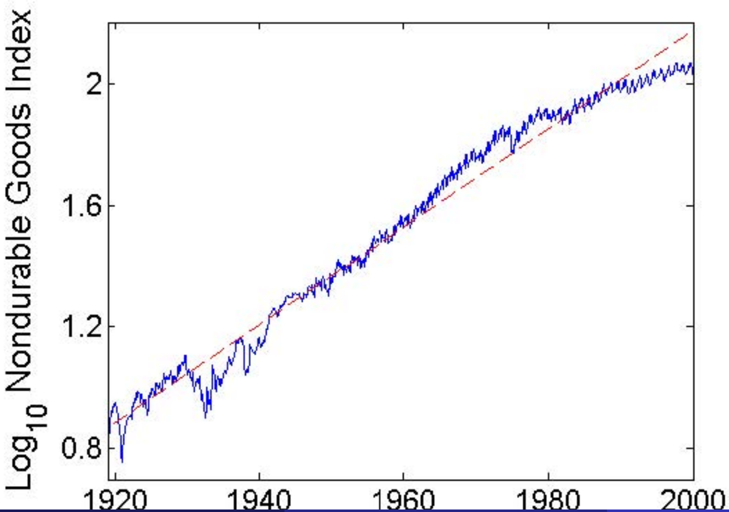
Functional data as a single long record

- Growth data contains growth measurement replications from independent inviduals.
- Many economic indicators are another type of functional data in form of a single long record.
- Example: Nondurable goods manufacturing index for the U.S.

U.S. nondurable goods manufacturing index



U.S. nondurable goods manufacturing index

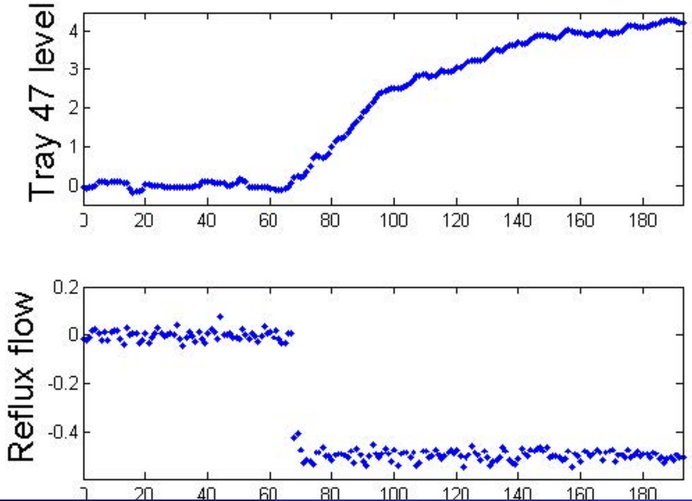


Multi-scale variation

These data, after transformation, have interesting variation on four different time scales:

- **Long term**: A remarkably linear trend with a slope of 1.6.
- **Medium term**: Multi-year changes due to the Great Depression (1929-1939), World War II (1939-1945), the Vietnam War (1955-1975), and over the last decade of the 20th century.
- **Short term**: Shocks like the stock market crash of 1928, the 1938 reduction of money supply and the end of the Vietnam War in 1975.
- **Seasonal effects**: Within-year effects that we will consider later, and that evolve smoothly from one year to the next.

An input/output system: oil refinery data

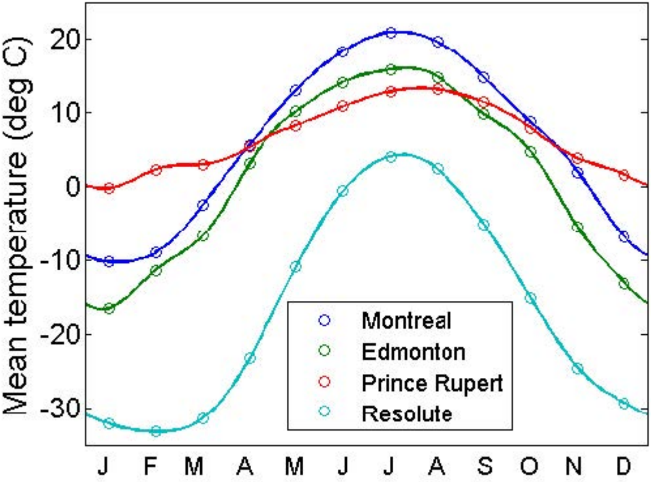


Oil refinery data

In an oil refinery in Texas,

- Top panel: amount of a petroleum product (measurement of tray level) in a distillation column or cracking tower.
- Bottom panel: flow of a vapor into the tray during an experiment.
- Production amount of the petroleum product reacts to the change in the flow of the vapor.
- **Analysis tool**: *concurrent functional linear model* with a functional response and a functional covariate.

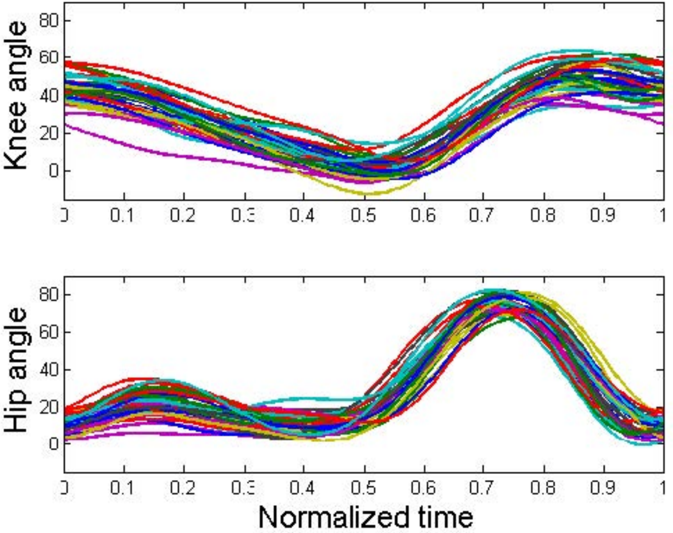
Temperatures at Canadian weather stations



Temperatures at Canadian weather stations

- Plots are mean monthly temperatures at four Canadian weather stations located respectively in Montreal, Edmonton, Prince Rupert, and Resolute.
- Without external *force*, one expects temperature to be sinusoidal.
- Question: how far are the temperatures at these stations from being sinusoidal?
- **Analysis tool**: *principal differential analysis*.

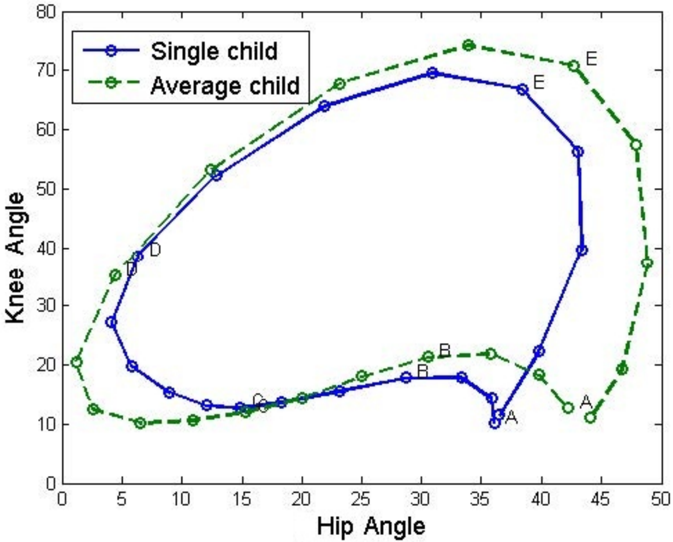
Multivariate functional data



Gait cycle data

- Study from the Motion Analysis Laboratory at Children's Hospital, San Diego.
- Plots are the angles in the sagittal plane formed by the hip and by the knee as 39 children go through a gait cycle.
- Interest in the interaction of the two periodic functions.
- **Analysis tool**: *functional canonical correlation analysis* and *creative graphical tools*.

One child's cycle vs. mean



Observations from the plot

- Points are equally spaced in time, starting from a heel strike (A) to the next heel strike (A) and forming a closed curve (ABCDEA).
- The spacings between the points show the angular velocity in each time segment.
- A cusp occurs at the heel strike.
- Differences between the child's and the average gaits:
  - In the C-D period when the hip is least bent, the child bends the hip more exaggeratively than the average.
  - In the E-A period when the hip is most bent, the child bends much less than the average.
- The child's knee angles are similar to the average.

Goals of functional data analysis

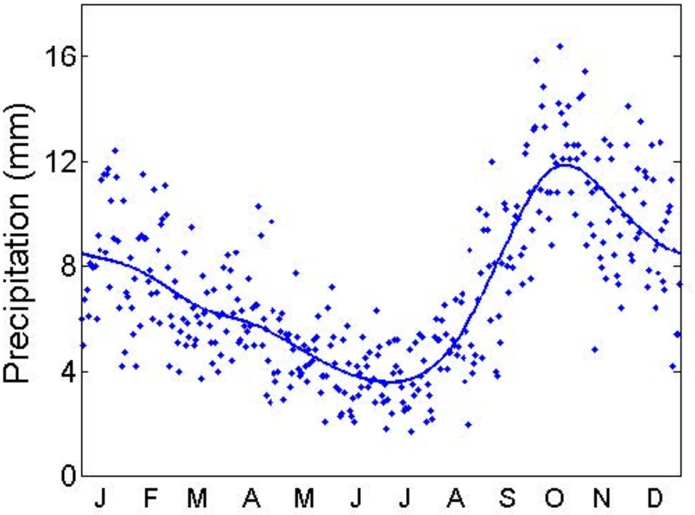
- to represent the data in ways that aid further analysis.
- to display the data so as to highlight various characteristics.
- to explain variation in an outcome or dependent variable by using input or independent variable information.
- to compare two or more sets of data with respect to certain types of variation, where two sets of data can contain different sets of replicates of the same functions, or different functions for a common set of replicates.

Smoothing and interpolation

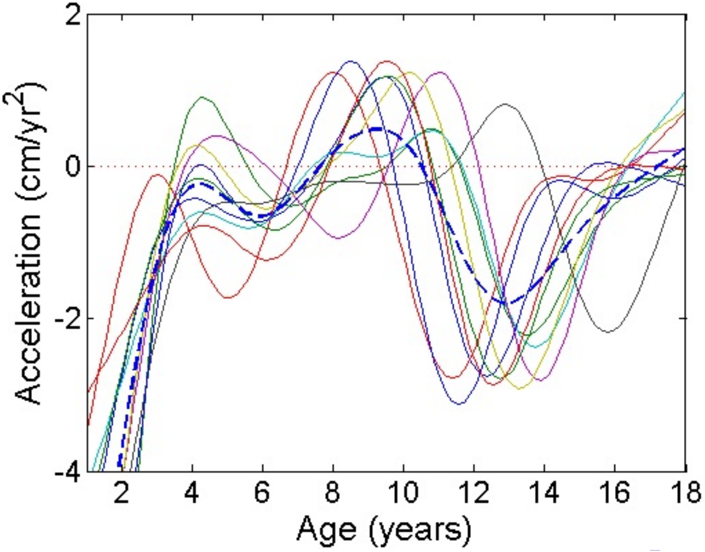
Functional data always come as discrete values. First step is to convert these values to a function.

- Interpolation: rarely used since observation errors are always expected.
- Basis function representation: Fourier basis system, spline basis system, wavelets, ...
- **Smoothing with roughness penalty**.

Rainfall at Prince Rupert



Data registration or feature alignment



The problem of phase variation

- Often important features in replicated curves do not occur at the same time (e.g., the pubertal growth spurt).
- *Phase variation* disrupts most obvious functional data analyses, which are designed for only *amplitude variation*.
- The mean curve here is a worthless summary of these growth acceleration curves.
- We must first align features, a process called *curve registration*.
- Registration separates phase and amplitude variation, which can then be studied independently, and also jointly.

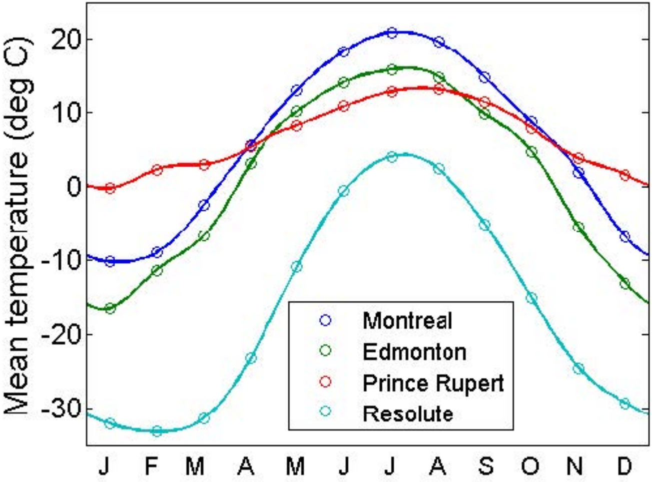
The sinusoidal component of temperature

- One expects temperature to be primarily sinusoidal in character, and certainly periodic over the annual cycle.
- There is much variation in level and some variation in phase.
- A model of the form

$$\mathbf{Temp}_i(t) \approx c_{i1} + c_{i2}\sin(\pi t/6) + c_{i3} cos(\pi t/6)$$

should do rather nicely for these data.

Temperatures at Canadian weather stations



- There are clear departures from sinusoidal or simple harmonic behavior.
- We could remove sinusoidal trend by regression, but let's use differentiation instead.
- We use $D^m x$ to refer to the mth derivative.
- We compute

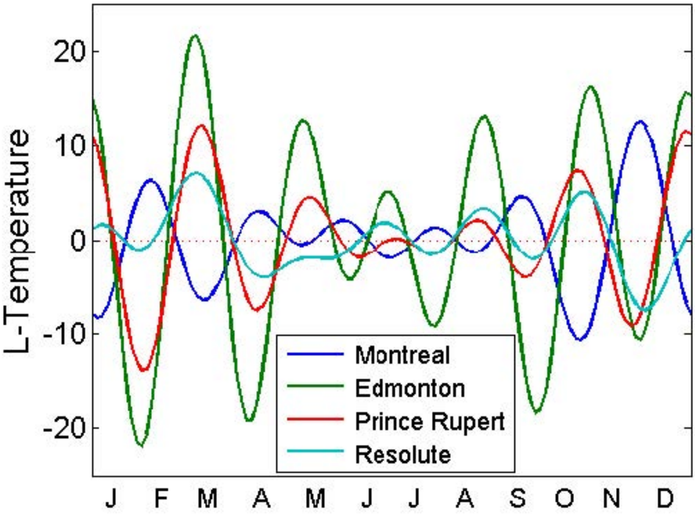
$$L\mathbf{Temp} \equiv (\pi/6)^2 D\mathbf{Temp} + D^3\mathbf{Temp}$$

  which will annihilate shifted sinusoids.
- L is a *linear differential operator*.
- We can define temperature as the solution to the nonhomogeneous differential equation

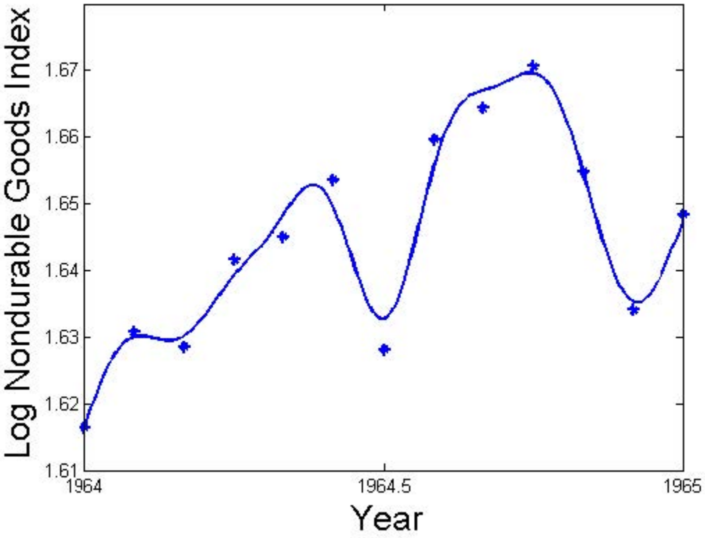
$$L\mathbf{Temp} = u$$

  where u is called a *forcing function*, and accounts for the non-sinusoidal effects.

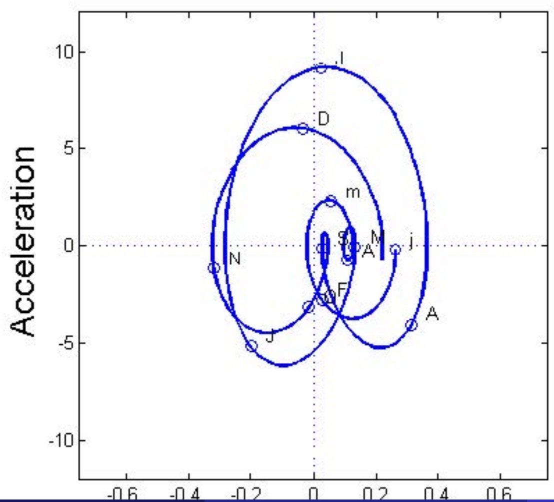
De-sined temperature



Seasonal trend in the 1964 goods index



Displaying seasonal dynamics: phase-plane plot



- Many types of functional data show strong *harmonic* variation.
- The acceleration or second derivative reflects *potential energy* in a mechanical system, like a pendulum or spring.
- The first derivative reflects its *kinetic energy*.
- A sinusoid is the prototype for such variation. Plotting its second derivative against first derivative produces a circle.
- The radius of the cycle is the total energy in the system, conserved as energy changes state.
- These ideas apply to most periodic phenomena.
- The phase-plane plot is a graphic version of a *differential equation*.

What makes FDA different?

- Unlike time series analyses, no assumptions of stationarity are made, and data are not sampled at equally spaced time points.
- Unlike most longitudinal data, a large number of time points are available, and the signal-to-noise ratio is medium to high.
- The data can support the accurate estimate of one or more derivatives, and these play several critical roles.
- Phase variation is recognized and separated from amplitude variation.
- Familiar multivariate methods have functional counterparts, and the smoothness of functional parameter estimates is explicitly controlled.
- Differential equations are new modeling tools.